IDIAP RESEARCH REPORT

TOKENWISE CONTRASTIVE SPEECH AND TEXT PRE-TRAINING FOR SPEECH EMOTION RECOGNITION

Eklavya Sarkar Neha Tarigopula

Idiap-RR-07-2025

AUGUST 2025

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Tokenwise Contrastive Speech and Text Pre-Training for Speech Emotion Recognition

Eklavya Sarkar
Idiap Research Institute
eklavya.sarkar@idiap.ch

Neha Tarigopula
Idiap Research Institute
neha.tarigopula@idiap.ch

Abstract

Human emotion recognition involves either decomposing audio signals to reflect the emotion or processing the corresponding text to extract the semantic meaning behind it. In this study, we explore the task of multi-modal emotion recognition by enriching acoustic representations with semantic meaning from the corresponding textual transcript. We use an pre-training strategy to learn the multi-modal representations via contrastive learning of token-by-token alignment of Whisper (speech) and BERT (text) representations using the LibriSpeech dataset. The aligned multi-modal features are then used for training an emotion classifier on IEMOCAP and EmoDB datasets. Despite the multi-modal representations outperforming the BERT-only uni-modal baselines, our results indicate a marginal underperformance compared to the Whisper-only uni-modal model, suggesting that leveraging additional textual information during pre-training might not necessarily improve representations for a downstream emotion recognition task.

1 Introduction

Speech emotion recognition (SER) is the task of automatically recognizing human emotions and affective states from oral speech, and has been studied rigorously in the last decades. It is an essential task in the human-computer interaction field, with applications in domains such as speech user interfaces, spoken language processing, and speech analysis for health (Schuller, 2018).

Typically, the approach for this task consists of finding the optimal audio feature representation for the given utterances, and using them as input to train a classifier. However, these methods focus only on the para-linguistic acoustic information of the utterance, such as the tonality, rhythm, intonation, prosody, loudness, and pitch, without any semantic knowledge of the spoken words.

To that end, some works have looked into leveraging the textual information in addition to the acoustic one for multimodal speech emotion recognition. Many of these look into combining speech and text features, extracted independently, into a joint representation using various fusion techniques. However, only a few works have attempted to jointly learn the alignment between speech and text at a deeper, token-level (Xu et al., 2019; Sunder et al., 2022). (Xu et al., 2019) learns the alignment via cross-attention between speech and text, to produce representations that are directly used for emotion classification. Whereas, (Sunder et al., 2022) uses a pre-training approach based on knowledge distillation to fuse speech and text representations for intent classification.

Typically in natural language processing (NLP), tokenization facilitates the generation of organized language representations, which aid in tasks such as semantic understanding and language modeling. It helps in determining the function and context of individual words, sub-words, or characters in a sentence. Whereas in speech, the modeling is at finer-auto frame level.

In this paper, we investigate whether learning a token-level alignment between speech and text can produce salient multi-modal representations enriched with semantic knowledge from textual information, that could potentially improve speech emotion recognition. Most of the recent success in end-to-end speech processing can be attributed to effective pre-training strategies to obtain robust speech representations. To that extent, we learn multi-model representations via a *pre-training* strategy on a large dataset of 960 hours of transcribed speech and apply them to a downstream task of emotion recognition. We hypothesize that using speech representations enriched with textual information will aid in emotion recognition.

The source code for the experiments can be found at: gitlab.idiap.ch/esarkar/course.ee608.project

2 Related Work

2.1 Unimodal approaches

The typical approaches for unimodal speech emotion recognition can be grouped into three types of feature extraction and modelling. The first consists of extracting hand-crafted acoustic features (El Ayadi et al., 2011) at a frame-level, compressing them into an utterance-level representation by computing statistical functionals or through bag-of-audio-words modeling, and then giving them as input to traditional classifiers such as Gaussian Mixture Models (Neiberg et al., 2006), Hidden Markov Models (Nwe et al., 2003; Schuller et al., 2003), Support Vector Machines (Mower et al., 2011), or Neural Networks (Stuhlsatz et al., 2011; Kim and Provost, 2013). The second consists of directly modeling at the utterance level with utterance level spectral features such as MFCCs and spectograms, or through end-to-end raw-waveform networks (Eyben et al., 2016; Neumann and Vu, 2017; Li et al., 2020; Kumawat and Routray, 2021). Finally, many approaches consist of leveraging pre-trained self-supervised learning (SSL) neural networks through a zero-shot classification, linear probing, or fine-tuning framework (Yang et al., 2021; Chen et al., 2022; Pepino et al., 2021).

2.2 Fusion methods for multimodal ER

Multimodal speech emotion recognition consists of extracting features from different domains, such as audio, text, or vision, and combining them through a fusion mechanism into a unified representation. These can be categorized into two main types: *model-agnostic fusion*, where the fusion is not directly dependent on a specific deep learning model, and *intermediate layer fusion*, which performs the fusion within a deep learning network.

The model-agnostic category consists of the following sub-categories of approaches:

- *Early fusion*: extracts feature vectors independently from the different modalities, and naively concatenates them together immediately after to give as a joint input to a single classifier (Lazaridou et al., 2015; Williams et al., 2018; Yoon et al., 2018). This method is predominant in the literature to successfully improve performance but does not perform any sort of alignment between the features of the different modalities.

- *Late fusion*: also referred to as decision-level fusion methods, independently extracts features and trains models for each modality, and then combines their prediction results through averaging, weighted sum, majority voting, or deep neural networks to obtain the output (Liu et al., 2014).

- *Hybrid fusion*: aims to obtain the optimal blend of features by strategically using early and late fusion strategies from different features to maximize the utilization of extracted emotional information.

The intermediate layer fusion consists of the following approaches:

- *Simple concatenation fusion*: similar to *early fusion*, but the concatenated features are high-level representations (embeddings) taken from a selected layer of a trained neural network instead of hand-crafted features.

- *Utterance-level interaction fusion*: explicitly models the features at the utterance-level across different modalities instead of simply concatenating them.

- *Fine-grained interaction fusion*: consists of cross-aligning the features of the different modalities at a token-level. Recent works have tried using the attention mechanism, which measures the similarity between features, to compute this alignment between extracted embeddings (Xu et al., 2019; Sunder et al., 2022).

In this work, we use the *fine-grained interaction fusion* to fuse our speech and text representations. However, unlike (Xu et al., 2019), we use a contrastive loss to align the text and speech sequences at a token level, and unlike (Sunder et al., 2022) we apply this to a speech emotion problem.

3 Methodology

The key idea behind our method is to distill knowledge from a text encoder to acoustic embeddings via token-by-token alignment of speech and text. An overview of the proposed framework is given in Figure 1, which consists of two key parts:

1. **Contextualization**: we use the speech representation of a given utterance to convert the non-contextual word embedding of the corresponding utterance's transcript into its contextualized form through a cross-modal attention mechanism.

2. **Alignment**: we implicitly inject fine-grained semantic knowledge from a 'ground truth' (i.e. contextualized) text-encoder representation into the speech representations through a contrastive loss.

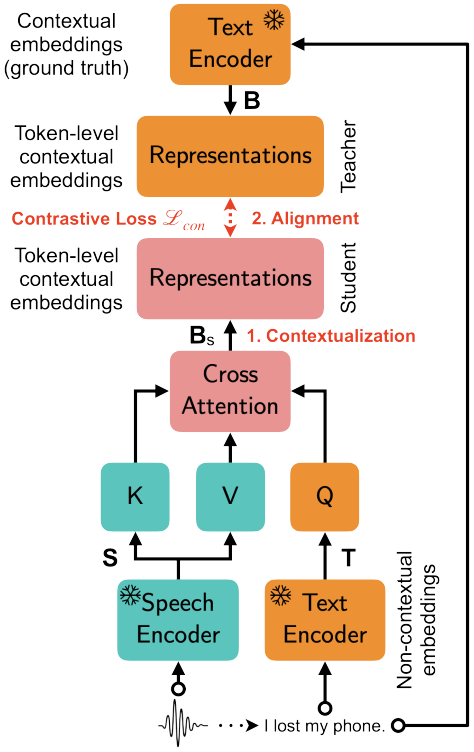

Figure 1: Architecture of the proposed model.

Intuitively this makes sense as the text encoder focuses on the individual token representations and the contextual information is induced from the speech representations, to learn a multi-model representation.

3.1 Speech encoders

We use the Whisper (Radford et al., 2023) model, a multi-lingually pre-trained model used for robust automatic speech recognition and translation, to extract the speech embeddings $\mathbf{S} \in \mathbb{R}^{N \times D}$, where D is the fixed encoder embedding dimension and N the variable number of frames. Specifically, we input the mel-filterbank features into the model and extract the representations from the encoder's last layer. We use both the base and large models for experiments.

3.2 Text encoders

We use the BERT (Devlin et al., 2019) model to encode the text transcripts of the corresponding speech utterance into text representations. Specifically, we pass the tokenized input sequence to the word embedding layer of BERT to obtain our non-contextual embeddings, $\mathbf{T} \in \mathbb{R}^{M \times D}$, where M is the number of tokens. These are used to contextualize the speech representations through cross-attention. The 'ground truth' teacher representations used for alignment step, denoted as $\mathbf{B} \in \mathbb{R}^{M \times D}$, are the contextual embeddings obtained from final encoder layer of BERT. We experiment with BERT trained on only English text data and also BERT trained on multi-lingual text data. We freeze both the speech and text encoders in our experiments.

3.3 Cross-attention contextualization

We convert the non-contextual text representations $\mathbf{T}$ into contextualized ones using the speech representations $\mathbf{S}$ and the cross-modal attention mechanism. Specifically, the speech representations are used as the keys and values, and the text non-contextual embeddings are used as the queries in the attention's dot-product computation. Thus, the representations $\mathbf{K} \in \mathbb{R}^{N \times D}$, $\mathbf{V} \in \mathbb{R}^{N \times D}$, and $\mathbf{Q} \in \mathbb{R}^{M \times D}$ are obtained as:

$$\mathbf{K} = \mathbf{S}\mathbf{W}_k$$
$$\mathbf{V} = \mathbf{S}\mathbf{W}_v$$
$$\mathbf{Q} = \mathbf{T}\mathbf{W}_q$$

Where $\mathbf{W}_k$, $\mathbf{W}_v$, and $\mathbf{W}_q \in \mathbb{R}^{D \times D}$ are learnable weights. The final output representations of the cross-attention, i.e. the contextual embeddings $\mathbf{B}_s \in \mathbb{R}^{M \times D}$ are obtained as:

$$\mathbf{B}_s = \text{softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$$

3.4 Contrastive alignment

The contextualized representation $\mathbf{B}_s$ can be aligned with the semantically rich 'ground truth' representation $\mathbf{B}$. This takes place at a token level given that both sequences have the same dimension M. To that end, we use a contrastive loss between pairs of token representations to inject fine-grained semantic knowledge. The cosine similarity between rows i and j of $\mathbf{B}$ and $\mathbf{B}_s$ is:

$$s_{ij} = \frac{\mathbf{B}_i \mathbf{B}_{sj}^T}{\tau \|\mathbf{B}_i\| \|\mathbf{B}_{sj}\|}$$

Where τ is a temperature parameter.

The contrastive loss is given as:

$$\mathcal{L}_{con} = -\frac{\tau}{2b} \sum_{i=1}^{b} \Theta$$

$$\Theta = \left[\log \frac{\exp(s_{ii})}{\Sigma_{j=1}^{b} \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\Sigma_{j=1}^{b} \exp(s_{ji})} \right]$$

It brings together representations of identical tokens (positive pairs), from the different modalities closer, while simultaneously distancing the representations of different tokens (negative pairs).

4 Experiments

4.1 Datasets

We pre-train the model on the 360, 500, and 960 hours train splits of the Librispeech dataset, consisting of read English from audiobooks, and sampled at 16 kHz. We report all our results on 960 hours of data. For the downstream evaluation, we used the EmoDB (Burkhardt et al., 2005) and IEMOCAP (Busso et al., 2008) datasets, which consist of utterances labeled with an emotion. They contain 7 and 5 emotion classes respectively, as detailed in Table 1, and consist of both scripted and improvised speech. We merge the *excited* and *happy* classes in the IEMOCAP dataset, following previous literature, and to keep the class distribution in the same magnitude.

Table 1: Dataset statistics.

	EmoDB	**IEMOCAP**
Language	German	English
Anger	127	1103
Happy	64	1636
Neutral	78	1708
Sad	52	1084
Disappointed	38	-
Fear	55	-
Bored	79	-
Total	**493**	**5531**

4.2 Baselines

We compare our extracted contextual and aligned representations with baselines of speech-only and text-only representations.

To that end, we use the Whisper encoder, as detailed in section 3.1 for our baseline. Additionally, based on the strong performance shown for emotion recognition on the SUPERB leaderboard (Yang et al., 2021), we also select WavLM, as another speech-only baseline. It is pre-trained on LibriSpeech 960h corpus with a masked speech denoising and prediction pre-text task. It contains 94.38M parameters and is leading the SUPERB challenge namely due to its 'full stack' speech representations. Due to a lack of consensus on the optimal layer for emotion recognition, we extract and use all thirteen encoder layers as independent features. We then only report the layer giving the best classification performance.

For our text-only representation, we simply use BERT's contextualized embeddings, obtained from the encoder's final layer, as described in section 3.2.

We input the selected speech models with the speech utterances to extract variable-length neural embeddings $\mathbf{S} \in \mathbb{R}^{N \times D}$ in a zero-shot framework, where D is the fixed encoder embedding dimension (512 and 768 for Whisper and WavLM respectively), and N the variable number of frames, contingent on the input utterance length. We then convert these embeddings into fixed-length statistical functionals $\boldsymbol{f}_{\mu\sigma} \in \mathbb{R}^{1 \times 2D}$ by computing and concatenating the first and second-order statistics across the frame axis on the extracted features. For BERT, we obtain an encoding representation $\mathbf{B} \in \mathbb{R}^{M \times D}$ of dimension $D = 768$ for each token, and compute the functionals with the identical methodology across the tokens axis.

4.3 Experimental Setup

Pre-training: We use the Librispeech 960h to pre-train the model to obtain representations. The model was trained for different batch sizes and we observed better results for the higher batch size of 64 using Adam optimizer with a learning rate of 1e-4. Figure 2 shows that the training converges over time. For the best model selection, we follow a k-NN validation strategy, where $k = 1$. We calculate the cosine distance between all the pairs of token representations obtained from the cross-attention and the teacher (BERT) and evaluate based on the percentage of matches (least distance between the corresponding pairs of tokens). Figure 3 shows the validation plots for different configurations of hyperparameters.

Downstream task: We split our emotion recognition dataset(s) of extracted features into *Train*, *Val*, and *Test* set following a 70:20:10 split protocol.

We train the model(s) on *Train* using Adam optimizer, learning rate η, and a cross-entropy loss for 30 epochs. We employ a dynamic learning rate scheduler to reduce the learning rate η when the selected optimization criterion, in this case *Val* UAR, shows no improvement after 10 epochs. The classifier is a simple feedforward network composed of

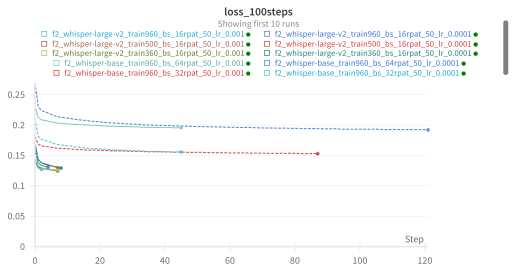

Figure 2: Convergence of the pre-training strategy.

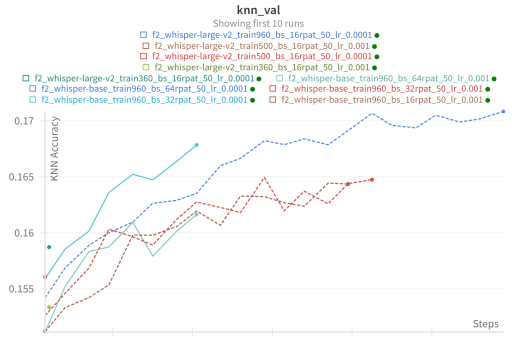

Figure 3: Validation plot for different configurations.

three blocks of [Linear, LayerNorm, ReLU] layers and a final linear layer.

To obtain the most robust classification, we employ a grid search methodology with the UAR score as the optimization criterion. We tune and evaluate the hyperparameters across the *Train* and *Val* sets respectively over the search space given in Table 2.

Table 2: Search space to find optimal hyperparameters.

Classifier	Hyperparameters	Search space
Baseline	Batch size	$2^{[2,10]}$
	η	1e-3, 5e-3
Proposed	Batch size	$2^{[2,5]}$
	η	1e-3, 1e-2
	Model	Base, large

We evaluate our multi-classification results using F1-Score, defined as the harmonic mean of the precision and recall scores, and Unweighted Average Recall (UAR) on *Test*.

Our experiments were conducted using a number of different tools, namely PyTorch (datasets and dataloaders), PyTorch Lightning (training and testing), Weights & Biases (logging and visualization), Hydra (experiment management), Optuna (hyperparameter sweeps), and Dask-Jobqueue (job launcher for Sun Grid Engine).

5 Results and Discussion

Table 3: Scores [%] on *Test*, using the optimal hyperparameters from the search space.

Dataset	Features	F1	UAR
EmoDB	BERT (Mono)	29.63	14.29
	BERT (Multi)	31.48	17.47
	Whisper	90.74	86.53
	WavLM	98.15	98.57
	Ours (Multi)	70.37	72.37
IEMOCAP	BERT (Mono)	68.95	69.13
	BERT (Multi)	66.25	66.51
	Whisper	74.55	75.9
	WavLM	90.74	86.53
	Ours (Mono)	73.65	74.32
	Ours (Multi)	74.73	75.28

Table 3 tabulates the results of emotion recognition using our representations from the pre-training against the other baselines. (Mono) refers to the representations from BERT trained on only English data and (Multi) refers to the representations from BERT trained on multi-lingual data.

We can observe from the results that the performance of text-only emotion recognition is far worse than the performance of speech-only and our fused speech-text. Furthermore, our fused speech-text representations surprisingly perform marginally worse than the speech-only representations. Given that in our experiments, we froze the whisper and BERT models for feature extraction, and updated only the cross-attention module, we conclude that the entire setup would benefit more from fine-tuning.

We also observe that the WavLM speech-only baseline outperforms all other models on both datasets, including Whisper. This follows existing literature which suggests that *audio*-based representations are more salient for the task of emotion recognition than pure *speech* representations.

The multilingual representations work better in almost all cases. For the text-only representations, we can see that the multilingual BERT works better on EmoDB than the monolingual version, but the inverse holds true for IEMOCAP. This follows our expectations as the EmoDB dataset is recorded in German, unlike IEMOCAP which is in English.

For the proposed framework, the model with the multilingual BERT also improves over monolingual one.

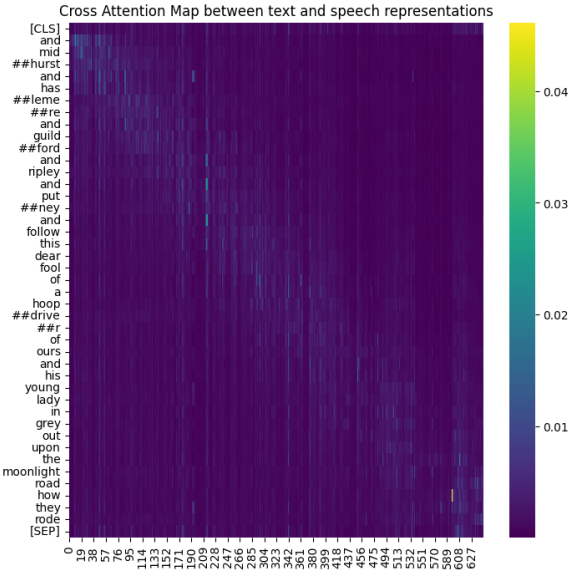

Figure 4: Cross-modal attention map between speech and text representations for one sentence from LibriSpeech dataset.

Figure 4 shows the attention map between speech and text representations for one utterance from the Lirbispeech dataset. We can observe that the relationship between the speech frames and text tokens is temporally coherent through the diagonal of the heatmap. Nonetheless, this monotonic alignment appears relatively subdued, corroborating with our prior results, indicating potential room for further adjustments to this framework to strengthen this alignment and improve the quality of the learnt representations.

6 Conclusion and Future Work

In this work, we investigated whether learning a contrastive multi-modal alignment between speech and text representations at a token-level could produce salient representations enriched with semantic knowledge from the textual information, and could thus improve on uni-modal representations for the downstream task for speech emotion recognition. To that end, we pre-trained multi-modal models in a proposed cross-attention and contrastive alignment framework, and then extracted the features from the learnt cross-attention to compare with those extracted from just the uni-modal encoder components. We found that although the multi-modal representations improve on the text-only representations, they perform marginally worse than the speech-only representations. This suggests that the additional representations in fact do not help to improve emotion recognition, and can actually decrease the performance.

For our future work, we would consider updating the text and speech encoders in the training phase to see if it helps to improve the downstream performance. We would also be curious to see if pre-training WavLM and BERT could yield a higher improvement over the corresponding speech-only baseline, when compared to Whisper and BERT. Finally, we would consider looking at the arousal information of the downstream utterances instead of just the emotion classes, as we suspect the scripted nature of the two downstream datasets could influence the results when compared to natural data.

Individual Contributions

Given the small scale of our team, consisting of just two students, both of us had a degree of contribution to all the different sections of this project. We estimate our individual contributions to be overall equal in this project, especially in the formulation of the research question and hypothesis, the discussion of the experimental results and conclusion, and debugging. A broad outline of our individual contributions is detailed below for clarity.

Eklavya

Eklavya implemented the main code repository, used for training the different models on the extracted features. To that end, his responsibilities included implementing feature standardization, classifier training, validation, and testing, as well as hyperparameter optimization over the defined search space. Additionally, he developed the PyTorch dataset/dataloaders to iterate over the EmoDB and IEMOCAP datasets and extracted the features for the speech-only baselines, namely Whisper and WavLM. Finally, he made significant contributions to the drafting the presentation and the report, as well as editing and reviewing the final versions.

Neha

Neha set up the dataloader for Librispeech and integrated it with the Whisper audio encoder. She also setup the framework of the proposed pre-training strategy and conducted the experiments on the Librispeech dataset across different hyperpameters. Furthermore, she was responsible for modifying

the dataloaders of IEMOCAP and EmoDB to include transcripts from the audio datasets and extracting the text-only BERT representations, the trained cross-attention features, and developing the attention map visualizations. Finally, she played a key role in the finalizing the project presentation and report.

References

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, volume 5, pages 1517–1520, Lisbon, Portugal. ISCA.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Yelin Kim and Emily Mower Provost. 2013. Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3677–3681.

Pooja Kumawat and Aurobinda Routray. 2021. Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition. In *Proc. Interspeech 2021*, pages 3410–3414.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Jeng-Lin Li, Tzu-Yun Huang, Chun-Min Chang, and Chi-Chun Lee. 2020. A waveform-feature dual branch acoustic embedding network for emotion recognition. *Frontiers in Computer Science*, 2.

Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. 2014. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 494–501, New York, NY, USA. Association for Computing Machinery.

Emily Mower, Maja J Matarić, and Shrikanth Narayanan. 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070.

Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. 2006. Emotion recognition in spontaneous speech using GMMs. In *Proc. Interspeech 2006*, pages paper 1581–Tue1A3O.5.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In *Proc. Interspeech 2017*, pages 1263–1267.

Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603–623.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*, pages 3400–3404.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

B. Schuller, G. Rigoll, and M. Lang. 2003. Hidden markov model-based speech emotion recognition. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages I–401.

Björn W. Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM*, 61(5):90–99.

André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. 2011. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691.

Vishal Sunder, Eric Fosler-Lussier, Samuel Thomas, Hong-Kwang Kuo, and Brian Kingsbury. 2022. Tokenwise Contrastive Pretraining for Finer Speech-to-BERT Alignment in End-to-End Speech-to-Intent Systems. In *Proc. Interspeech 2022*, pages 2683–2687.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2019. Learning Alignment for Multimodal Emotion Recognition from Speech. In *Proc. Interspeech 2019*, pages 3569–3573.

Shuwen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118.